Bayesian Sparse Factor Analysis of Genetic Covariance Matrices

Daniel E. Runcie*, Sayan Mukherjee†

Abstract

Quantitative genetic studies that model complex, multivariate phenotypes are important for both evolutionary prediction and accurate livestock selection. For example, changes in gene expression can provide insight into developmental and physiological mechanisms that link genotype and phenotype. However, classical analytical techniques are poorly suited to quantitative genetic studies of gene expression where the number of traits assayed per individual can reach many thousand. Here, we derive a Bayesian sparse factor model for estimating the genetic covariance matrix (G-matrix) of high-dimensional traits, such as gene expression. The key idea of our model is that we need only consider G-matrices that are biologically plausible. An organism's entire phenotype is the result of developmental processes that are modular and have limited complexity. This implies that the G-matrix will be highly structured. In particular, we assume that a limited number of intermediate traits (or factors, e.g., variations in development or physiology) control the variation in the high-dimensional phenotype, and that each of these intermediate traits is sparse – affecting only a few measured traits. The advantages of this approach are three-fold. First, sparse factors are interpretable and provide biological insight into mechanisms underlying the genetic architecture. Second, enforcing sparsity helps prevent sampling errors from swamping out the true signal in high-dimensional data. Third, our Bayesian analysis automatically provides credible intervals for the

*Department of Biology, Duke University, Durham, NC 27708

†Departments of Statistical Science, Computer Science, and Mathematics, Duke University, Durham, NC 27708

heritability of measured and intermediate traits. We demonstrate the advantages of our model on simulated data and re-analyze gene expression data from a natural population of *Drosophila melanogaster*.

Keywords G-matrix, factor model, sparsity, Bayesian inference, animal model

1 Introduction

Understanding evolutionary change requires knowledge of phenotypes in a population, as well as the genetic architecture underlying phenotypic variation (Houle, 2010). It is well known in quantitative genetics that multiple correlated traits need to be modeled jointly to avoid unexpected or counterintuitive evolutionary outcomes (Walsh and Blows, 2009). While most evolutionary studies have focused on external traits, such as morphology or coloration, there is an increasing effort to collect more comprehensive phenotypic information. Traits such as behavior, development or physiology, and molecular signatures such as cellular metabolism are also important for determining fitness and driving evolution. Including a comprehensive set of phenotypes is likely to enrich evolutionary studies and lead to more accurate livestock and crop selection. However, these traits tend to be tightly associated at several levels: morphology is the output of development; behavior drives and can be driven by physiology; cellular metabolism powers growth. Applying the tools of quantitative genetics to these high-dimensional and highly correlated datasets introduces considerable analytical and computational challenges. In this paper we formulate a modeling framework to address these challenges.

The most basic quantitative genetics analysis partitions total phenotypic variation into genetic and environmental components (Lynch and Walsh, 1998). The central quantity of interest in many analyses is the matrix of additive genetic variances and covariances among traits, called the G-matrix. The G-matrix encodes information about evolutionary potential in a set of traits. This is highlighted by the Breeder's or Lande equation (Lande, 1979), which states that when selection is applied to a set of traits, the expected response is the G-matrix times the selection gradient – the G-matrix rotates and scales the selection gradient and can potentially shift the direction of change in each trait. An eigendecomposition of the G-matrix provides important insight into this process. Selection gradients aligned with eigenvectors corresponding to large eigenvalues are expected to have a strong response. In

particular, the eigenvector with the largest eigenvalue has been called the line of genetic least resistance and is thought to bias evolutionary change (Schluter, 1996). Similarly, if the G-matrix is singular then certain selection directions (in particular, selection vectors in the nullspace) will be ineffective even if all the traits selected on are variable. Such constraints are reduced when the G-matrix is highly modular. Modularity in the G-matrix – groups of traits that are genetically correlated but uncorrelated with other traits – captures both pleiotropy as well as traits that evolve independently (Cheverud, 1996). Estimating the G-matrix is thus a key step in many quantitative genetic analyses.

There has been increasing interest in including gene expression as a trait in evolutionary analyses (Ayroles *et al.*, 2009; Mcgraw *et al.*, 2011; Gibson and Weir, 2005). Genome-wide gene expression assays target thousands of traits simultaneously and provide a means to monitor cellular and developmental traits that would otherwise be difficult to measure. Databases of gene function and gene expression signatures of cellular perturbations, stresses or disease states are continually growing, and have been successful in turning gene expression measurements into biologically insights. Using expression as a trait is also appealing since variation in expression is expected to have a simpler genetic basis than fitness or yield, yet can have a large effect on fitness. Genetic analyses of gene expression phenotypes can also identify sets of genetically co-variable transcripts and infer novel molecular networks.

The challenge in scaling methods developed in quantitative genetics to hundreds or thousands of traits is primarily methodological. The number of parameters required to infer the G-matrix grows as $p(p+1)/2$, where p is the number of traits. The situation is compounded if we are also required to model environmental variation or measurement error (Kirkpatrick and Meyer, 2004). The huge number of parameters coupled with modest numbers of individuals in evolutionary studies can lead to instability in parameter estimates. Extracting biological insight from a large matrix is also a challenge. Individual genetic variances or covariances can be highly misleading if not considered with respect to the overall structure of the G-matrix – even for traits with strong genetic variances or covariances (Walsh and Blows, 2009; Hine and Blows, 2006).

Our objective in this paper is to develop a model for estimating G-matrices that is scalable to large numbers of traits and is applicable to a wide range of datasets, including both experimental crosses and pedigreed populations. Previous methods for estimating additive genetic variation and covariation can be categorized as fol-

lows: (1) pairwise estimates of genetic covariation followed by clustering (Ayroles *et al.*, 2009; Stone and Ayroles, 2009), (2) methods based on moments estimators (Hine and Blows, 2006; Mcgraw *et al.*, 2011), and (3) methods based on mixed effects models (Henderson, 1984; Kruuk, 2004; Kirkpatrick and Meyer, 2004; de Los Campos and Gianola, 2007). The shortcoming of the pairwise approach (Ayroles *et al.*, 2009) is that simply collecting pairwise covariance estimates will not in general result in a proper covariance matrix – the inferred G-matrix may not be positive (semi)definite. Methods based on moments estimators (Hine and Blows, 2006; Mcgraw *et al.*, 2011) are generally not flexible enough to model more involved experimental designs such as wild populations or large breeding programs. Estimators based on the "Animal Model" (Henderson, 1984) address the above problems by fitting a linear mixed model (LMM) using pedigree information to partition the observed phenotypic variance of a trait into various genetic and environmental components. The LMM can be applied to a much broader range of experimental designs and studies (Kruuk, 2004), and produces estimates in the parameter space. However, these methods are computationally costly for high-dimensional data. Dimension-reduction approaches such as principle components analysis on the high-dimensional trait vector followed by univariate analysis (Biswas *et al.*, 2008) can reduce computation demands, but are problematic if there is significant environmentally induced covariation among traits (Meyer and Kirkpatrick, 2010; Mcgraw *et al.*, 2011). Efficient mixed model approaches for moderate-dimension data include using Restrictred Maximum Likelihood (REML) to fit the eigenvectors corresponding to the largest eigenvalues of the G-matrix (Kirkpatrick and Meyer, 2004), and a Bayesian approach constraining the G-matrix to take the form of a factor model with a limited number of latent traits (de Los Campos and Gianola, 2007). However, neither of these LMM approaches (Kirkpatrick and Meyer, 2004; de Los Campos and Gianola, 2007) as formulated will scale to high-dimensional trait data.

The key idea that allows us to infer G-matrices for large numbers of traits is that the matrix of additive genetic variation will likely be both sparse and modular. Here, sparsity means that many of the values in the G-matrix (or a factorization of the matrix) will be zero, and modular means that groups of traits will covary together. Our *a priori* assumption is that the G-matrix is composed of only a few modules (factors), and that few (sparse) traits will have significant effects in each module. We therefore constrain the class of covariance matrices we can estimate, a necessary procedure for inference of covariance matrices given high-dimensional data

(Bickel and Levina, 2008b,a; el Karoui, 2008; Meyer and Kirkpatrick, 2010; Carvalho *et al.*, 2008; Hahn *et al.*, 2013).

The biological argument behind this prior assumption starts with the observation that the phenotypes that we measure in an organism are formed by a shared developmental process and this developmental process has limited complexity. For gene expression, regulatory networks and functional pathways control gene expression and variation in gene expression can be often linked to genetic variation in pathways (Xiong *et al.*, 2012; de la Cruz *et al.*, 2010). For a given dataset, we make two assumptions about these pathways: (1) a limited number of pathways are relevant for trait variation and (2) each pathway affects a limited number of genes. There is support and evidence for these modeling assumptions in the quantitative genetics literature as G-matrices tend to be highly structured (Walsh and Blows, 2009) and the majority of genetic variation is contained in a few dimensions regardless of the number of traits studied (Ayroles *et al.*, 2009; Mcgraw *et al.*, 2011).

In this paper we provide a Bayesian sparse factor model for inferring G-matrices from pedigree information for hundreds or thousands of traits. This model is an extension of the classic multivariate animal model, and so is highly flexible. We demonstrate the advantages of the model on simulated data and re-analyze gene expression data from a published study on *Drosophila melanogaster* (Ayroles *et al.*, 2009). Although high-dimensional sparse models have been widely used in genetic association studies (Cantor *et al.*, 2010; Engelhardt and Stephens, 2010; Stegle *et al.*, 2010; Parts *et al.*, 2011; Zhou and Stephens, 2012) to our knowledge, sparsity has not been applied to estimating a G-matrix.

2 Methods

We will specify a Bayesian factor model that encodes the two main biological assumptions we make on the G-matrix: sparsity in the number of factors comprising the matrix, and sparsity in each factor – each factor is comprised of a few components. This factor model is designed to address the high-dimensional setting where hundreds or thousands of traits are simultaneously examined. The sparsity assumption is the key feature in our model that allows us to scale stable and accurate inference to a very large number of traits. For high-dimensional models sparsity helps prevent sampling errors from swamping out the true signal in data leading to stable parame-

ter estimates. In our model, sparsity implies that each underlying trait will effect few of the observable phenotypes and as a result many of the parameters in the model will be (near) zero.

2.1 Model:

We derive the Bayesian sparse factor model as an extension to the classic multivariate animal model. For a single trait the following linear mixed effects model is used to explain phenotypic variation (Henderson, 1984):

$$\mathbf{y}_i = \mathbf{X}\mathbf{b}_i + \mathbf{Z}\mathbf{u}_i + \mathbf{e}_i \tag{1}$$

where $\mathbf{y}_i$ is the vector of phenotype measurements for trait i on n individuals, $\mathbf{b}$ is a vector of fixed effects and environmental covariates such as sex or age, with design matrix $\mathbf{X}$, $\mathbf{u}_i$ is the random vector of additive genetic effects with covariance $\sigma_i^2\mathbf{A}$, where $\mathbf{A}$ is the known additive relationship matrix among the individuals and $\mathbf{Z}$ relates the random effects to the observations, and $\mathbf{e}_i$ are error which are assumed to be independent of the additive genetic effects.

In going from one trait to p traits we can stack the vector $\mathbf{y}_i$ for each trait in (1) to a $n \times p$ matrix $\mathbf{Y}$ specified by the following linear mixed effects model:

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{Z}\mathbf{U} + \mathbf{E} \tag{2}$$

where the random terms $\mathbf{U} = [\mathbf{u}_1 \dots \mathbf{u}_p]$ and $\mathbf{E} = [\mathbf{e}_1 \dots \mathbf{e}_p]$ are drawn from matrix normal distributions

$$\mathbf{U} \sim \mathrm{MN}_{r,p}(\mathbf{0}; \mathbf{A}, \mathbf{G}), \quad \mathbf{E} \sim \mathrm{MN}_{n,p}(\mathbf{0}; \mathbf{I}_n, \mathbf{R}), \tag{3}$$

where $\mathbf{0}$ is the $n \times p$ matrix of zeros, $\mathbf{G}$ and $\mathbf{R}$ are the $p \times p$ matrices modeling genetic and residual covariances among traits, and $\mathbf{A}$ is the known additive relationship matrix among the individuals, with rank $r \leq n$. The matrix normal distribution is defined as

$$p(\mathbf{V} \mid \mathbf{M}, \mathbf{\Omega}, \mathbf{\Sigma}) = \frac{\exp\left(-\frac{1}{2}\mathrm{tr}[\mathbf{\Omega}^{-1/2}(\mathbf{V}-\mathbf{M})^T\mathbf{\Sigma}^{-1}(\mathbf{V}-\mathbf{M})]\right)}{(2\pi)^{np/2}|\mathbf{\Omega}|^{n/2}|\mathbf{\Sigma}|^{p/2}}.$$

We model the genetic and residual covariance matrices with a factor structure

$$\begin{aligned} \mathbf{G} &= \mathbf{\Lambda}_u\mathbf{\Sigma}_u\mathbf{\Lambda}_u^T + \mathbf{\Psi}_u, \\ \mathbf{R} &= \mathbf{\Lambda}_e\mathbf{\Sigma}_e\mathbf{\Lambda}_e^T + \mathbf{\Psi}_e, \end{aligned} \tag{4}$$

where $\mathbf{\Lambda}_u$ and $\mathbf{\Lambda}_e$ are $p \times k_u$ and $p \times k_e$ factor loading matrices, $\mathbf{\Sigma}_u$ and $\mathbf{\Sigma}_e$ are $k_u \times k_u$ and $k_e \times k_e$ diagonal matrices, and $\mathbf{\Psi}_e$ and $\mathbf{\Psi}_u$ are $p \times p$ diagonal matrices. We can now specify $\mathbf{U}$ and $\mathbf{E}$ via the following hierarchical model

$$\begin{aligned} \mathbf{U} = \mathbf{F}_u \mathbf{\Lambda}_u^T + \mathbf{\Delta}, \quad & \mathbf{E} = \mathbf{F}_e \mathbf{\Lambda}_e^T + \mathbf{\Xi} \\ \mathbf{F}_u \sim \mathrm{MN}_{r,k_u}(\mathbf{0}; \mathbf{A}, \mathbf{\Sigma}_u), \quad & \mathbf{F}_e \sim \mathrm{MN}_{n,k_e}(\mathbf{0}; \mathbf{I}_n, \mathbf{\Sigma}_e) \\ \mathbf{\Delta} \sim \mathrm{MN}_{r,p}(\mathbf{0}; \mathbf{A}, \mathbf{\Psi}_u), \quad & \mathbf{\Xi} \sim \mathrm{MN}_{n,p}(\mathbf{0}; \mathbf{I}_n, \mathbf{\Psi}_e). \end{aligned} \tag{5}$$

In the above model $\mathbf{F}_u$ and $\mathbf{F}_e$ are latent traits corresponding to additive genetic effects and residuals, respectively, and $\mathbf{\Sigma}_u$ and $\mathbf{\Sigma}_e$ model the covariances of these latent effects.

In (5), we assume that the underlying genetic and residual effects are unique and fit the factors $\mathbf{\Lambda}_u$ and $\mathbf{\Lambda}_e$ separately. This need not be the case since some factors may influence both $\mathbf{U}$ and $\mathbf{E}$.

To generalize to factors driving both genetic and residual effects we rewrite (2) as

$$\mathbf{Y} = \mathbf{XB} + \mathbf{F}\mathbf{\Lambda}^T + \mathbf{Z}\mathbf{\Delta} + \mathbf{\Xi} \tag{6}$$

where $\mathbf{F} = \mathbf{Z}\mathbf{F}_u^* + \mathbf{F}_e^*$ is a single set of latent traits with $\mathbf{F}_u^* = [\mathbf{F}_u \ \mathbf{0}_{r,k_e}]$ and $\mathbf{F}_e^* = [\mathbf{0}_{n,k_u} \ \mathbf{F}_e]$, and $\mathbf{\Lambda} = [\mathbf{\Lambda}_u \ \mathbf{\Lambda}_e]$ with $k = k_u + k_e$. We can specify $\mathbf{F}$ via the following matrix normal distributions

$$\mathbf{F}_u^* \sim \mathrm{MN}_{r,k}\left(\mathbf{0}, \mathbf{A}, \begin{bmatrix} \mathbf{\Sigma}_u & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}\right), \quad \mathbf{F}_e^* \sim \mathrm{MN}_{n,k}\left(\mathbf{0}, \mathbf{I}_n, \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{\Sigma}_e \end{bmatrix}\right), \tag{7}$$

where $\mathbf{\Sigma}_u = \mathrm{Diag}(\sigma_{u_j}^2)$ and $\mathbf{\Sigma}_e = \mathrm{Diag}(\sigma_{e_j}^2)$.

In addition to inference of the latent traits $\mathbf{F}$ themselves, we would like to infer the heritability of each latent trait. By (7) the columns of $\mathbf{F}$ are independent, and by marginalizing over $\mathbf{F}_u$ and $\mathbf{F}_e$ we obtain the distribution of each factor as

$$\mathbf{f}_j \sim \mathrm{N}_n(\mathbf{0}, \sigma_{u_j}^2 \mathbf{Z}\mathbf{A}\mathbf{Z}^T + \sigma_{e_j}^2 \mathbf{I}_n). \tag{8}$$

Heritability of the latent trait $\mathbf{f}_j$, is the ratio of its additive genetic variance to its total (phenotypic) variance: $h_j^2 = \frac{\sigma_{u_j}^2}{\sigma_{u_j}^2 + \sigma_{e_j}^2}$. Reparameterizing the above distribution leads to

$$(\sigma_{u_j}^2 + \sigma_{e_j}^2)^{-1/2} \mathbf{f}_j \sim \mathrm{N}_n(0, h_j^2 \mathbf{Z}\mathbf{A}\mathbf{Z}^\mathbf{T} + (1 - h_j^2)\mathbf{I}_n). \tag{9}$$

This makes heritability explicit as an inference parameter. The relevance of this is that the latent traits themselves are likely to have both genetic and non-genetic influences, like any other trait. Without loss of generality we can scale Λ_j by the phenotypic standard deviation of factor j, $\sqrt{\sigma^2_{u_j} + \sigma^2_{e_j}}$, and renormalize $\sigma^2_{u_j} + \sigma^2_{e_j} = 1$. In this formulation, the key matrices $\mathbf{G}$ and $\mathbf{R}$ can be recovered as:

$$\begin{aligned} \mathbf{G} &= \mathbf{\Lambda}\,\mathrm{diag}(h_i^2)\,\mathbf{\Lambda}^T + \mathbf{\Psi}_u \\ \mathbf{R} &= \mathbf{\Lambda}\,\mathrm{diag}(1 - h_i^2)\,\mathbf{\Lambda}^T + \mathbf{\Psi}_e. \end{aligned} \tag{10}$$

2.2 Prior specification:

The model specified in (6) is identical to the standard multivariate mixed model (2). It is clear in modeling high-dimensional data (Meyer and Kirkpatrick, 2010) that that stability and accuracy of parameter estimates is a serious problem and some prior specification or penalty/regularization is required for robust estimates – that is constraints on $\mathbf{G}$ and $\mathbf{R}$ are required. We impose constraints on $\mathbf{G}$ and $\mathbf{R}$ through highly informative priors on Λ. Our priors are based on two key biological assumptions based on the idea that the genetic and residual covariances arise from variation in underlying developmental processes which are driven by gene networks or metabolic pathways. This implies:
(1) The biological system has limited complexity – a limited number of pathways are relevant for trait variation, $k \ll p$. For the model this means that the number of factors is low.
(2) Each underlying developmental pathway affects a limited number of traits. For the model this means the factor loadings are sparse.

We formalize the above assumptions by priors on $\mathbf{\Lambda}$ that impose sparsity (Bhattacharya and Dunson, 2011). Sparsity in $\mathbf{\Lambda}$ will impose constraints on the genetic and residual covariance matrices. Sparsity on the loadings, assumption (2), is imposed by a heavy tailed distribution on elements λ_{im} of the factor loadings which favors small values while allowing a few large values. Limits on the number of factors, assumption (1), is imposed by a prior that shrinks the magnitude of successive factors. The prior is specified as a hierarchical distribution on each element λ_{im} of

Λ:

$$\begin{aligned}
\lambda_{im} \mid \phi_{im}, \tau_m &\sim \mathrm{N}\left(0, \phi_{im}^{-1}\tau_m^{-1}\right) \\
\phi_{im} &\sim \mathrm{Ga}(\nu/2, \nu/2), \\
\tau_m &= \prod_{l=1}^{m} \delta_l, \\
\delta_1 &\sim \mathrm{Ga}(a_1, b_1), \\
\delta_l &\sim \mathrm{Ga}(a_2, b_2) \text{ for } l = 2, ..., k.
\end{aligned} \tag{11}$$

Shrinkage on the number of factors is imposed by the parameter τ_m which increasingly shrinks all elements of high-index columns of $\mathbf{\Lambda}$. This shrinkage is induced by the the stochastically increasing product of the sequence $\{\delta_l\}$. Sparsity in the factor loadings is controlled by ϕ_{im} which governs the precisions of λ_{im}. Conditional on τ_m, marginalizing out the ϕ_{im} leads to a t-distribution for each λ_{im} with ν degrees of freedom – this is the heavy tail that imposes sparsity on the factor loadings.

For heritability of each latent trait we set as a prior a discrete set of values in the unit interval. This was done for computational efficiency.

$$\pi(h_i^2 = l/n_h) = 1/n_h, \text{ where } l = 0 \ldots (n_h - 1) \tag{12}$$

In principle, we could place a prior on the interval $[0, 1]$, but such a prior would not be conjugate, and so coding a MCMC sampler slightly more difficult.

We place gamma priors on each of the inverse variances on the diagonals of Ψ_u and Ψ_e. Priors on each element of $\mathbf{B}$ are normal distributions with very large ($> 10^6$) variances.

2.3 Implementation:

Inference in the above model uses an adaptive Gibbs sampler for which we provide detailed steps in the appendix. The code has been implemented in Matlab® and can be found at the website (http://stat.duke.edu/∼sayan/quantmod.html).

3 Results

3.1 Simulation example:

To test the performance of our model, we generated 50 simulated datasets from a pedigree using the *R* (R Core Team, 2012) package *pedantics* (Morrissey, 2010). The *gryphons* pedigree in this package is designed to be relevant to power and sensitivity analyses for quantitative genetic studies of natural populations (Morrissey, 2010). We selected 148 individuals from last cohort of the pedigree and their 39 mothers as our sample population. All individuals had at least one relative in the sample so that the pedigree was moderately informative. We simulated phenotypic data using the function *phensim*. For each simulation, we constructed sparse, modular matrices $\mathbf{G}$ and $\mathbf{R}$ for $p = 100$ traits. We began by generating a set of $k = 8$ sparse factors with non-zero loadings for 10-50 traits. We then assigned positive heritabilities to $k_u = 5$ of these factors, each drawn from an independent Beta$(3, 2)$ distribution. We set $\mathbf{\Psi}_u = 0.2 \times \mathbf{I}_p$ and $\mathbf{\Psi}_e = 0.5 \times \mathbf{I}_p$, and calculated $\mathbf{G}$ and $\mathbf{R}$ using equation (10). In these simulations, the narrow-sense heritabilities (h^2) of the 100 measured traits ranged from $\sim 0.02 - 0.8$, with the majority < 0.2. We calculated the genetic relationship matrix $\mathbf{A}$ as 2-times the matrix of kinship coefficients calculated from the full *gryphons* pedigree using the *kinship* package (Atkinson and Therneau, 2012).

We compared our Bayesian sparse factor model to a standard pairwise mixed model analysis. We set the prior hyperparameters: $\nu = 3, a_1 = 2, b_1 = 1/20, a_2 = 3, b_2 = 1$. For each simulation, we ran our Gibbs sampler for 12,000 iterations, discarded the first 2,000 samples as a burn-in period, and collected 1000 posterior samples with a thinning rate of 10. We then used the program WOMBAT (Meyer, 2007) to fit each of the $p(p + 1)/2$ genetic covariances in the G-matrix separately using REML. Our Bayesian sparse factor model provided more accurate estimates of these genetic covariances (Figure 1). On average, the mean square error (MSE) of the covariance estimates, a measure of absolute accuracy, was 33% lower with the factor model than with the pairwise analysis (Figure 2A). The improvement approached 80% in cases when the data were more informative, but the quality of the fits converged in simulations in which neither model estimated the genetic covariances particularly well. The correlation between estimated and actual genetic covariances, a measure of the accuracy of relative magnitudes of covariances among

genes, was consistently higher with the factor model, with an average improvement of $> 40\%$ (Figure 2B). The factor model's covariance estimates were positively biased, while the pairwise REML estimates were not (Figure 2C). h^2 estimates from the factor model had similar MSE to the REML estimates, but were also consistently positively biased.

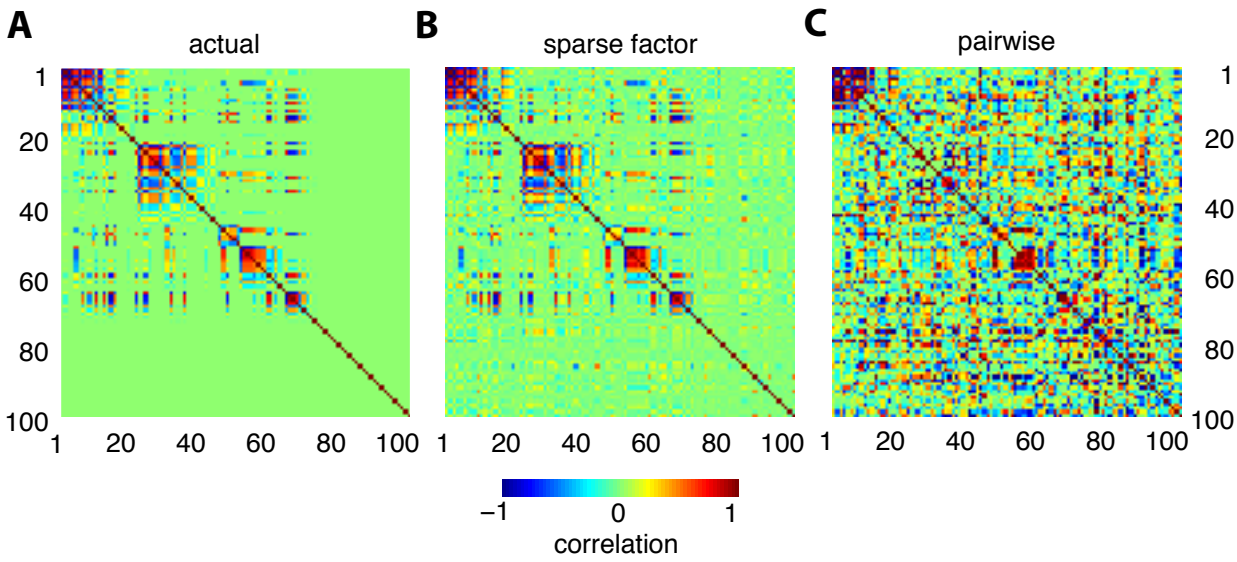

Figure 1: **The Bayesian genetic sparse factor model accurately estimates the G-matrix from a pedigree.** **A**. Matrix of simulated genetic correlations. A G-matrix was simulated for 100 traits with five factors (see Methods). The G-matrix was normalized to have unit genetic variances on the diagonal for visual clarity. Phenotypic data from 187 individuals simulated from the *gryphons* pedigree (Morrissey, 2010) was generated given this G-matrix. **B**. Matrix of posterior mean genetic correlations estimated from these simulated individuals by the Bayesian genetic sparse factor model for 100 traits. **C**. Matrix of genetic correlations estimated by REML from two-trait animal model analyses run using WOMBAT (Meyer, 2007). All panels share the same color scale.

Importantly, the Bayesian sparse factor model produced estimates of the G-matrix that were positive definite and accurately fit the underlying latent trait structure. In our simulations with 100 traits, element-wise estimates of the G-matrix by

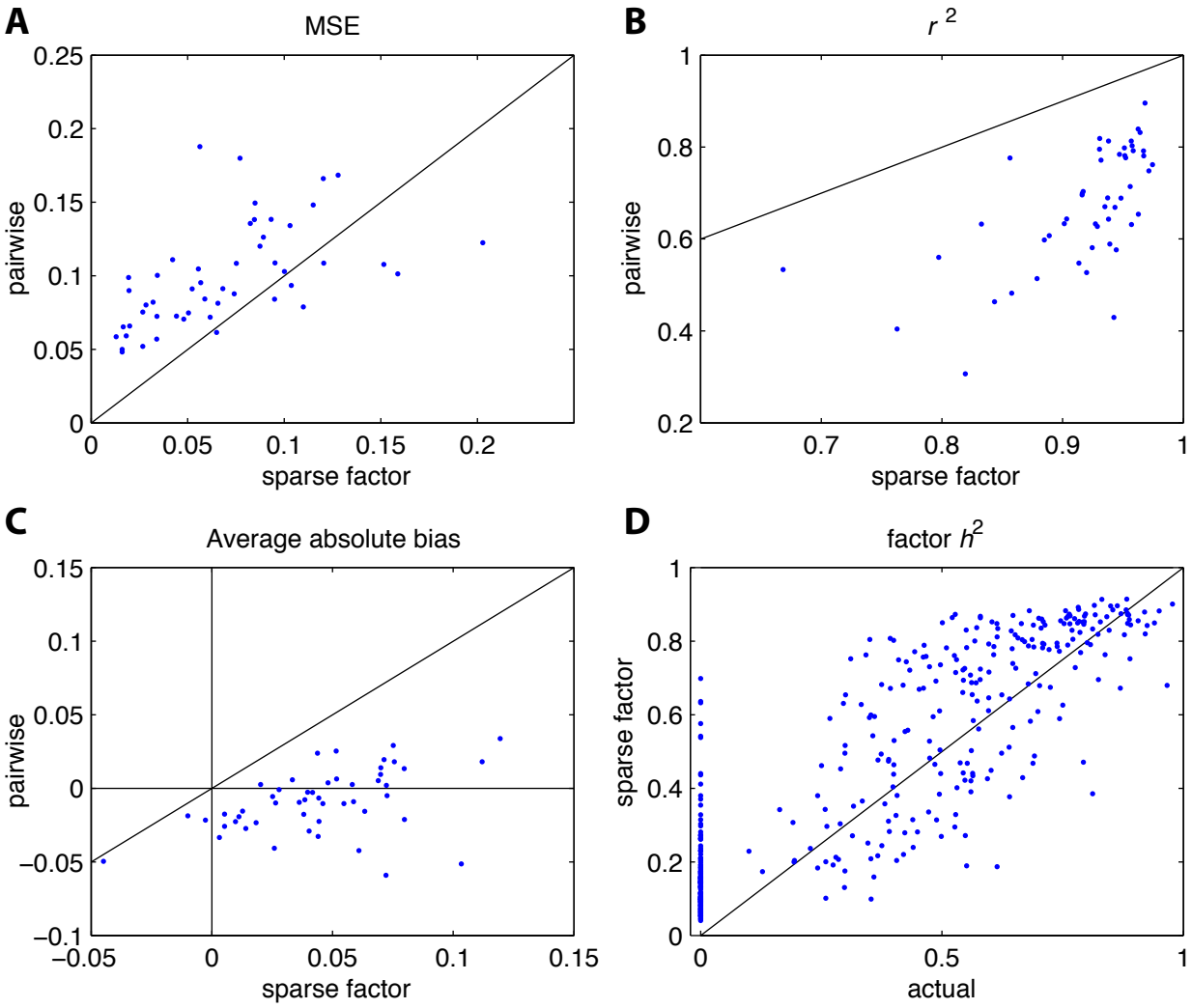

Figure 2: **Joint estimation across all traits leads to quantitative improvements in parameter estimates.** Pairwise genetic correlations among 100 traits from simulations of 187 individuals were estimated by two methods: 1) The Bayesian sparse factor model for 100 traits; 2) Repeated (pairwise) REML using WOMBAT. Panels **A-C** compare the accuracy and bias of the sparse factor model estimates to the pairwise REML estimates across 50 simulations. Each point represents the results of one simulation. **A**. Mean squared error of genetic covariances. **B**. Squared Pearson correlations of estimated and actual genetic covariances. **C**. Average absolute bias in genetic covariances. **D**. Accuracy of latent trait heritabilities. For each simulation, fitted latent traits were matched to simulated factors by calculating the (absolute) correlation between the columns of $\mathbf{\Lambda}$ and the true trait loadings on each factor. In this plot, known heritabilities of each of the eight simulated factors (x-axis) are compared to the heritability estimates (y-axis) of the most correlated latent traits in that simulation. All 50 simulations are combined. The diagonal line in each plot shows $y = x$.

REML had 40-48 negative eigenvalues, and thus could not easily be inverted in calculations of genetic selection gradients (Rausher, 1992). In all 50 simulations, the columns of $\mathbf{\Lambda}$ included each of the 8 simulated latent traits. And, because of the sparsity prior, no *a posteriori* rotation of the factors was needed for interpretability (e.g., Meyer 2009, Figure 3). Furthermore, the estimates of the heritabilities of latent traits were accurate ($r = 0.87$ over all 50 simulations, after matching up each simulated latent traits with the most correlated column of $\mathbf{\Lambda}$, Figure 2D).

3.2 Gene expression example:

We downloaded gene expression profiles of 40 wild-derived lines of *Drosophila melanogaster* from ArrayExpress (accession: E-MEXP-1594, Ayroles *et al.* 2009) and used our Bayesian sparse factor model to infer an among-line gene expression covariance matrix for a subset of the genes. We first normalized the processed gene expression data to correspond to the the analyses of the earlier paper and then selected the 414 genes that Ayroles *et al.* (2009) identified as having a plausible among-line covariance with competitive fitness. We also downloaded competitive fitness data for each line from the Drosophila Genetic Reference Panel (DGRP) website (http://dgrp.gnets.ncsu.edu/) so that we could estimate the among line covariance between each gene and this phenotypic trait.

In this dataset, two biological replicates of male and female fly collections from each line were analyzed for whole-animal RNA expression. The competitive fitness measurements were means of 20 competitive trials done with sets of flies from these same lines, but not the same flies used in the gene expression analysis. Therefore, gene expression values for the samples measured for competitive fitness and competitive fitness values for the samples measured for gene expression were treated as missing data (see Appendix). We used our model to estimate the G-matrix of the genes (in this case, the covariance of line effects). Following the analyses of Ayroles *et al.* (2009), we included a fixed effect of sex, and independent random effects of the sex:line interaction for each gene. No sex or sex:line effects were fit for competitive fitness itself as this value is measured at the level of the line, not individual flies.

We set the prior hyperparameters as above, and ran our Gibbs sampler for 40,000 iterations, discarded the first 20,000 samples as a burn-in period, and collected 1,000 posterior samples of all parameters with a thinning rate of 20. Our estimate of the G-matrix was qualitatively similar to the original estimate (Figure 4A, and compare

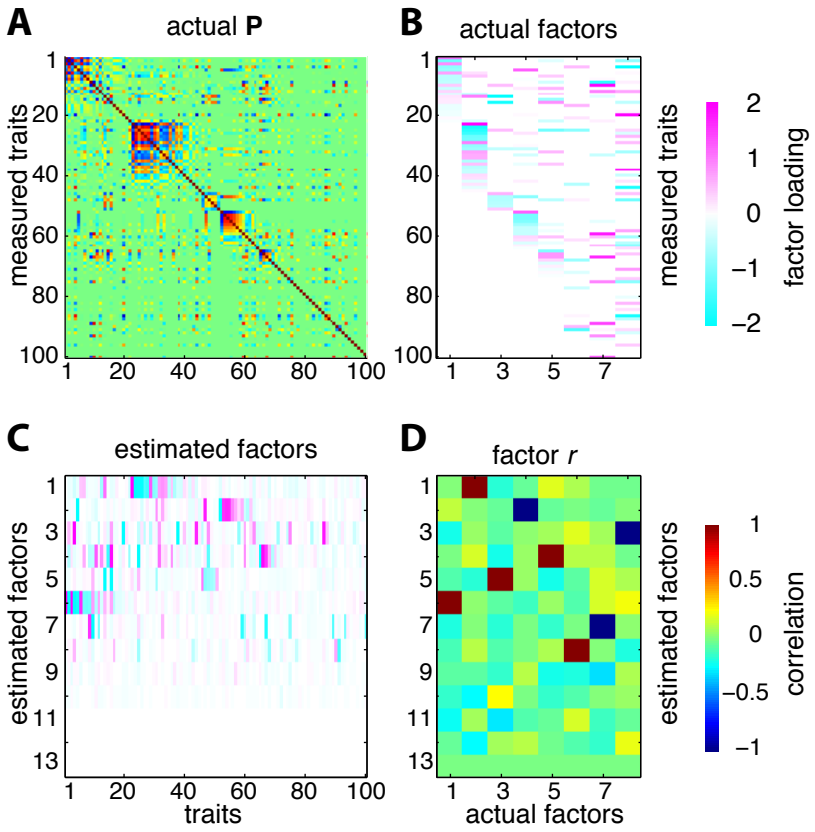

Figure 3: **The Bayesian genetic sparse factor model recovers the underlying latent factors in simulated data.** These panels represent the same simulated dataset as in Figure 1. **A**. Matrix of phenotypic correlations ($\mathbf{P} = \mathbf{G}+\mathbf{R}$, normalized to have unit variances). **B**. Simulated latent traits. Each column represents the loadings of the 100 measured traits on one of the eight simulated latent traits. Traits 1-5 were assigned non-zero heritablities. **C**. Posterior mean estimates of the latent traits under the Bayesian genetic sparse factor model. The model selected 13 factors, but only eight had large loadings for any of the measured traits. **D**. Pearson correlations (r) between the loadings of the eight simulated traits, and posterior means of the 13 factors. All eight simulated latent traits were recovered in the estimated factors with $|r| > 0.96$. Panels **A** and **D** share the same color scale, as do panels **B** and **C**. The latter color scale is truncated for clarity.

to Figure 7a in Ayroles *et al.* (2009)). However, the estimate from our factor model was positive definite, while the original estimate was not since it was calculated in a pairwise fashion. Nevertheless, estimates of the broad-sense heritability of each gene were similar ($r = 0.74$).

Using the Modulated Modularity Clustering (MMC) algorithm (Stone and Ayroles, 2009), Ayroles *et al.* (2009) identified 20 modules of genetically correlated transcripts *post-hoc*. In our factor model, modules are estimated simultaneously with the G-matrix itself. Each factor (column of $\mathbf{\Lambda}$) represents a sparse set of genes that are highly correlated in their expression, possibly due to common regulation by some latent developmental trait. Our model identified 27 such latent factors (Figure 4B). Of these factors, 13-16 of them were consistently identified ($r > 0.95$) across 3 parallel chains of the Gibbs sampler, and most of the rest were minor, each accounting for less that 1% of the variance in any gene.Many factors were similar, but not identical to the modules identified by MMC (Figure 4B). Some of the factors were nearly one-to-one matches to modules (e.g., factor 10 with module 8, and factor 14 with module 12). However, others merged together two or more modules (e.g., factor 1 with modules 7 and 9, and factor 2 with modules 4, 13, 16-20). And some entire modules were part of two or more factors (e.g., module 17 was included in factors 2 and 4, and module 18 was included in factors 2 and 16).

One reason for the discrepancy between our factor model and the MMC results is that our model allows each gene to belong to more than one of the latent traits. A second difference is that our model infers factors at the level of phenotypic variation, rather than the among-line covariances. The broad-sense heritability (H^2) of the latent traits (factors) ranged from 0.03 to 0.90 (Figure 4B). The majority of the factors, had intermediate H^2, between 0.1 and 0.65, but 5 were largely genetic with $H^2 > 0.75$.

Although a functional analysis would be more powerful if more genes were studied simultaneously, the latent traits defined by the modules do capture intriguing biological relations: using the Database for Annotation, Visualization and Integrated Discovery (DAVID) v6.7 (Huang *et al.*, 2009a,b), several of the factors were individually enriched (within this set of 414 genes) for proteins related to processes such as: defense and immunity, nervous system function, odorant binding, transcription and cuticle formation. Similar molecular functions were identified among the modules identified by Ayroles *et al.* (2009). These authors highlighted modules 7-9 in particular – modules 7 and 9 contained largely female-biased genes, while module 8

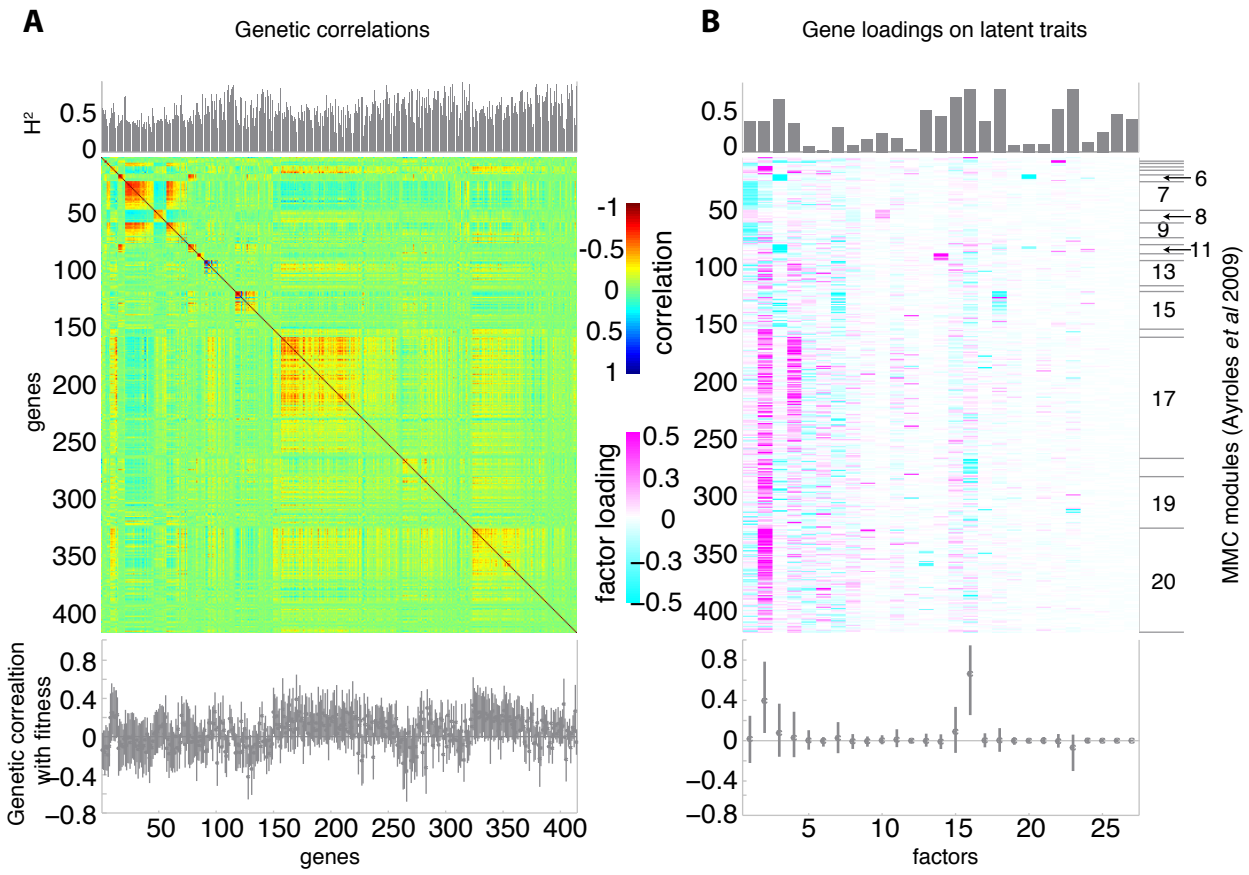

Figure 4: **Among-line covariance of gene expression and competitive fitness in *Drosophila* is modular.** Gene expression (414 genes) and competitive fitness data for 40 wild-derived lines of *Drosophila melanogaster* were downloaded from ArrayExpress (accession: E-MEXP-1594) (Ayroles *et al.*, 2009) and http://dgrp.gnets.ncsu.edu/. **A**. Genetic (among-line) architecture of gene expression traits. The three panels show: i) Posterior mean broad-sense heritabilities (H^2) for the 414 genes, ii) Posterior mean genetic correlations among these genes, and iii) Posterior means and 95% highest posterior density (HPD) intervals around estimates of genetic correlations between each gene and competitive fitness. For comparison, see Figure 7a of (Ayroles *et al.*, 2009). **B**. Latent trait structure of gene expression covariances. The three panels show: i) Posterior mean H^2 for each estimated latent trait, ii) Posterior mean gene loadings on each latent trait, and iii) Posterior means and 95% (HPD) intervals around estimates of genetic correlations between each latent trait and competitive fitness. The right-axis of panel **B**. groups genes into modules inferred using Modulated Modularity Clustering (Stone and Ayroles, 2009; Ayroles *et al.*, 2009).

contained male-biased genes – and hypothesized that negative genetic correlations between genes in the female- and male-biased modules could act to maintain genetic variation in fitness. In our model, modules 7 and 9 are largely grouped into factor 1, which also weakly includes the genes of module 8. However, factor 10 is the major contributor to genetic variation in the genes of module 8.

Finally, by adding a competitive fitness as a 415th trait in the analysis, we could estimate the among-line correlation between the expression of each gene and this fitness-related trait (Figure 4A). Many ($60/414 \sim 15\%$) of the 95% highest posterior density (HPD) intervals on the among-line correlations did not included zero, although most of these correlations were low (for 85% of genes, $|r| < 0.25$) with a few as large as $|r| \sim 0.45$. We also estimated the genetic correlation between competitive fitness and each of the latent traits defined by the 27 factors (Figure 4B). Here, most correlations were nearly zero. However, the genetic correlations between competitive fitness and factors 2 and 16 were large and highly significant, suggesting potentially interesting genetic relationships between these underlying traits and fitness.

4 Discussion

The Bayesian sparse factor model performs well on both simulated and real data, and thus opens the possibility of incorporating diverse and highly complex traits into evolutionary genetic studies and breeding programs. Gene expression traits in particular provide a way to measure under-appreciated molecular and developmental traits that may be important for evolution, and technologies exist to measure these traits on very large scales. Our model can also be applied to other molecular traits (e.g., metabolites or protein concentrations), high dimensional morphological traits (e.g., outlines of surfaces from geometric morphometrics), or gene-environment interactions (e.g., the same trait measured in multiple environments).

4.1 Scalability of the method:

The key advantage of the Bayesian sparse factor model over existing methods is its ability to provide robust estimates of covariance parameters for datasets with large numbers of traits. In this study, we demonstrated very high performance of the model for 100 simulated traits, and robust results on real data with 415. Similar factor models (without the genetic component) have been applied to gene

expression datasets with thousands of traits (Bhattacharya and Dunson, 2011), and we expect the genetic model to perform similarly. The main limitation will be computational time, which scales roughly linearly with the number of traits analyzed (assuming the number of important factors grows more slowly). Parallel computing techniques may speed up analyses in cases of very large numbers of traits.

The main reason that our model scales well in this way is that under our prior, each factor is sparse. Experience with factor models in fields such as gene expression analysis, economics, finance, and social sciences (Fan *et al.*, 2011), as well as with genetic association studies (e.g., Engelhardt and Stephens 2010; Stegle *et al.* 2010; Parts *et al.* 2011) demonstrates that sparsity (or shrinkage) is necessary to perform robust inference on high-dimensional data (Bickel and Levina, 2008b,a; el Karoui, 2008; Meyer and Kirkpatrick, 2010). Otherwise, sampling variability can overwhelm any true signals, leading to unstable estimates. Here, we used the t-distribution as a shrinkage prior, following (Bhattacharya and Dunson, 2011), but many other choices are possible (Armagan *et al.*, 2011).

4.2 Applications to evolutionary quantitive genetics:

The G-matrix features prominently in the theory of evolutionary quantitative genetics, and its estimation has been a prominent goal of many experimental and observational studies (Walsh and Blows, 2009). Since our model is built on the standard mixed effect model framework, it is flexible and can be applied to many experimental designs or studies. And since our model is Bayesian and naturally produces estimates within the parameter space, posterior samples from the Gibbs sampler provide convenient credible intervals for the G-matrix itself and many evolutionarily important parameters, such as trait-specific heritabilities or individual breeding values (Sorensen and Gianola, 2010).

An important use of G-matrices is to predict the response of a set of traits to selection (Lande, 1979). Applying Robertson's 2nd theorem of natural selection, the response in $\bar{\mathbf{y}}$ will equal the additive genetic covariance between the vector of traits and fitness ($\Delta\bar{\mathbf{y}} = \sigma_A(\mathbf{y}, \bar{w})$) (Rausher, 1992; Walsh and Blows, 2009). This quantity can be estimated directly from our model if fitness is included as the $p^* = (p+1)$th trait:

$$\Delta\bar{\mathbf{y}} = \Lambda_{u_{/p^*}} \Lambda_{u_{p^*}}^T,$$

where $\Lambda_{u_{/p^*}}$ contains all rows of Λ_u except the row for fitness, and $\Lambda_{u_{p^*}}$ contains

only the row of Λ_u corresponding to fitness. Similarly, the quantity $1 - \Psi_{u_{p^*}}/\mathbf{G}_{p^*,p^*}$ equals the percentage of genetic variation in fitness accounted for by variation in the measured traits (Walsh and Blows, 2009), which is useful for identifying other traits that might be relevant for fitness.

On the other hand, our model is not well suited to estimating the dimensionality of the G-matrix. A low-rank G-matrix means that there are absolute genetic constraints on evolution (Lande, 1979). Several methods provide statistical tests for the rank of the G-matrix (e.g., Hine and Blows 2006; Kirkpatrick and Meyer 2004; Mezey and Houle 2005). We use a prior that shrinks the magnitudes of higher index factors to provide robust estimates of the largest factors. This will likely have a side-effect of underestimating the total number of factors. However, absolute constraints appear rare (Houle, 2010), and the dimensions of the G-matrix with the most variation are likely those with the greatest effect on evolution in natural populations (Schluter, 1996; Kirkpatrick, 2009). Our model should estimate these dimensions well. From a practical standpoint, pre-selecting the number of factors has plagued other reduced-rank estimators of the G-matrix (e.g., Kirkpatrick and Meyer 2004; Hine and Blows 2006; Meyer 2009). Our prior is based on an infinite factor model (Bhattacharya and Dunson, 2011), and so no *a priori* decision is needed. Instead, the parameters of the prior distribution become important modeling decisions. In our experience, a relatively diffuse prior on δ_l with $a_2 = 3$ tends to work well.

4.3 Biological interpretation of factors:

Genetic modules are sets of traits likely to evolve together. By assuming that the developmental process is modular, we can model each latent trait as affecting a limited number of phenotypic traits. Other techniques for identifying genetic modules include the MMC algorithm (Stone and Ayroles, 2009; Ayroles *et al.*, 2009), and spectral decomposition which treats each major eigenvector as an estimate of such an underlying module (e.g., Mcgraw *et al.* 2011). The former technique constraints each trait to belong to only one module, while the biological interpretation of the latter is unclear because of the mathematical constraint that the eigenvectors be orthogonal (Hansen and Houle, 2008). In classic factor models (such as proposed by Meyer (2009), or de Los Campos and Gianola (2007)), the factors are not identifiable (Meyer, 2009), and so the identity of the underlying modules is unclear.

Under our sparsity prior, the factors are identifiable up to a sign-flip (each factor can be multiplied by -1 without affecting its probability under the model). However, in simulations and with the *Drosophila* gene expression data, our Gibbs sampler (see the Appendix) chooses a single sign of each factor for long stretches of each chain. Also, independent chains identify the same dominant factors. The ordering of the factors is also constrained in our model. The prior on δ_h makes factors with large loadings on large numbers of traits increasingly improbable for higher-index factors. Under vague priors on δ_h, the order of similarly indexed factors can be different among MCMC chains. In general, the order of the factors is not of great biological interest. However, as more high-dimensional datasets are created and studied, more informative priors on δ_h may be justified and will likely reduce this problem.

A unique feature of our model is the fact that we estimate genetic and environmental factors jointly, instead of separately as in classic multilevel factor models (e.g., Goldstein 2010). If each factor represents a true latent trait (e.g., variation in a developmental process), it is reasonable to decompose variation in this trait into genetic and environmental components. We directly estimate the heritability of the traits underlying each factor, and therefore can use our model to predict the evolution of these latent traits.

4.4 Extensions:

Our model is built on the classic mixed effect model common in quantitative genetics (Henderson, 1984). It is therefore straightforward to extend to models with additional fixed or random effects (e.g., dominance or epistatic effects) for each trait. The update equation for h_j^2 in the Gibbs sampler described in the Appendix however does not allow additional random effects in the model for the latent factors themselves ($\mathbf{f}_j$ in equation (8)), although other formulations are possible. A second extension relates to the case when the relationship matrix among individuals ($\mathbf{A}$) is unknown. Here, relationship estimates from genotype data can be easily incorporated. As such, our model is related to a recently proposed sparse factor model for genetic associations with intermediate phenotypes (Parts *et al.*, 2011). These authors introduced prior information on genetic modules from gene function and pathway databases which could be incorporated in our model in a similar way.

5 Conclusions

The Bayesian sparse factor model for genetic analysis that we propose provides a novel approach to genetic estimation with high-dimensional traits. We anticipate that incorporating many diverse phenotypes into genetic studies will provide powerful insights into evolutionary processes. The use of highly-informative but biologically grounded priors is necessary for making inferences on high-dimensional data, and can help identify developmental mechanisms underlying phenotypic variation in populations.

6 Appendix

6.1 Posterior sampling:

We estimate the posterior distribution of the Bayesian genetic sparse factor model with an adaptive Gibbs sampler based on the procedure proposed by Bhattacharya and Dunson (2011). The value k^* at which columns in $\mathbf{\Lambda}$ are truncated is set using an adaptive procedure (Bhattacharya and Dunson, 2011). Given a truncation point, the sampler iterates through the following steps:

1. If missing observations are present, values are drawn independently from univariate normal distributions parameterized by the current values of all other parameters:

$$\pi(y_{ij} \mid -) \sim \mathrm{N}\left(\mathbf{x}^{(j)}\mathbf{b}_i + \mathbf{f}^{(j)}\lambda_i + \mathbf{z}^{(j)}\delta_i, (\sigma_i^{-2})^{-1}\right)$$

   where y_{ij} is the imputed phenotype value for the i-th trait in individual j. The three components of the mean are: $\mathbf{x}^{(j)}$ the row vector of fixed effect covariates for individual j times $\mathbf{b}_i$, the ith column of the fixed effect coefficient matrix; $\mathbf{f}^{(j)}$, the row vector of factor scores on the k^* factors for individual j times λ_i, the row of the factor loading matrix for trait i; and $\mathbf{z}^{(j)}$, the row vector of the random (genetic) effect incidence matrix for individual j times δ_i, the vector of residual genetic effects for trait i not accounted for by the k^* factors. Finally, σ_i^{-2} is the residual precision of trait i. All missing data can be drawn in a single block update.

2. The fixed effect coefficient matrix $\mathbf{B}$, the truncated factor loading matrix $\mathbf{\Lambda}_{k^*}$ and the residual genetic effects matrix $\mathbf{\Delta}$ can be stacked into a single matrix, and then its columns factor into independent multivariate normal conditional posteriors:

$$\pi\left(\left[\begin{array}{c}\mathbf{b}_i\\ \lambda_i\\ \delta_i\end{array}\right]\Big|-\right) \sim \mathrm{N}\left(\mathbf{C}^{-1}\mathbf{W}^T\sigma_i^2\mathbf{y}_i, \mathbf{C}^{-1}\right),$$

where $\mathbf{W}$ and $\mathbf{C}$ are defined as:

$$\begin{aligned}\mathbf{W} &= [\mathbf{X}\ \mathbf{F}\ \mathbf{Z}]\\ \mathbf{C} &= \left[\begin{array}{ccc}\mathbf{0} & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \psi_{u_{ii}}^{-2}\mathbf{A}^{-1} & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathrm{Diag}(\phi_{jl}\tau_l)\end{array}\right] + \sigma_i^{-2}\mathbf{W}\mathbf{W}^T.\end{aligned}$$

3. The conditional posterior of the factor scores $\mathbf{F}$ is a matrix variate normal distribution:

$$\pi\left(\mathbf{F} \mid -\right) \sim \mathrm{MN}_{n,k^*}\left(\mathbf{C}^{-1}\left(\tilde{\mathbf{Y}}\mathbf{\Psi}_e^{-1}\mathbf{\Lambda}_{\mathbf{k}^*} + \mathbf{Z}\mathbf{F}_u\mathrm{Diag}(1-h_i^2)^{-1}\right), \mathbf{C}^{-1}\right)$$

where $\mathbf{C}$ is:

$$\mathbf{C} = \mathbf{\Lambda}_{k^*}^T\mathbf{\Psi}_e^{-1}\mathbf{\Lambda}_{k^*} + \mathrm{Diag}(1-h_i^2)^{-1}$$

and $\tilde{\mathbf{Y}}$ is:

$$\tilde{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\mathbf{B} - \mathbf{Z}\mathbf{\Delta}.$$

4. The conditional posterior of the genetic effects on the factors, $\mathbf{F}_u$ factors into independent multivariate normals for each factor $\mathbf{f}_{u_m}, m = 1 \ldots k^*$ st $h_m^2 \neq 0$:

$$\pi\left(\mathbf{f}_{u_m} \mid -\right) \sim \mathrm{MN}\left(\mathbf{C}^{-1}(1-h_m^2)^{-1}\mathbf{Z}\mathbf{F}_m, \mathbf{C}^{-1}\right)$$

where $\mathbf{C}$ is:

$$\mathbf{C} = (1-h_m^2)^{-1}\mathbf{Z}\mathbf{Z}^T + (h_m^2)^{-1}\mathbf{A}^{-1}.$$

5. The conditional posterior for each of the latent factor heritabilities $h^2_m, m = 1 \dots k^*$ is calculated by integrating out $\mathbf{F}_u$ and summing over all possibilities of h^2_m, since the prior on this parameter is discrete:

$$\pi\left(h^2_m = h^2 \mid -\right) = \frac{\mathrm{N}\left(\mathbf{F}_m \mid \mathbf{0}, h^2\mathbf{ZAZ}^T + (1-h^2)\mathbf{I}_n\right)\pi(h^2_m = h^2)}{\sum_{l=1}^{n_h} \mathrm{N}\left(\mathbf{F}_m \mid \mathbf{0}, h^2_l\mathbf{ZAZ}^T + (1-h^2_l)\mathbf{I}_n\right)\pi(h^2_m = h^2_l)}$$

where $\mathrm{N}(\mathbf{x} \mid \mu, \Sigma)$ is the multivariate normal density with mean μ and variance Σ, evaluated at $\mathbf{x}$, $h^2_l = l/h_h$, and in general, $\pi(h^2_m = h^2_k) = 1/n_h$. Given this conditional posterior, h^2_m is sampled from a multinomial distribution.

6. The conditional posterior of the trait-factor loading variance ϕ_{ih} for trait i on factor h is:

$$\pi(\phi_{ih} \mid -) \sim \mathrm{Ga}\left(\frac{\nu+1}{2}, \frac{\nu+\lambda^2_{ih}}{2}\right).$$

7. The conditional posterior of $\delta_m, m = 1 \dots k^*$ is as follows. For δ_1:

$$\pi(\delta_1 \mid -) \sim \mathrm{Ga}\left(a_1 + \frac{pk^*}{2}, b_1 + \frac{1}{2}\sum_{l=1}^{k^*}\tau^{(1)}_l \sum_{j=1}^{p}\phi_{jl}\lambda^2_{jl}\right)$$

and for $\delta_h, h \geq 2$:

$$\pi(\delta_h \mid -) \sim \mathrm{Ga}\left(a_2 + \frac{p}{2}(k^* - h + 1), b_2 + \frac{1}{2}\sum_{l=h}^{k^*}\tau^{(h)}_l \sum_{j=1}^{p}\phi_{jl}\lambda^2_{jl}\right)$$

where $\tau^{(h)}_l = \prod_{t=1, t\neq h}^{l} \delta_t$ for $h = 1 \dots k^*$.

8. The conditional posteriors for the precision of the residual genetic effects of trait i, $\psi_{u_{ii}}$, is:

$$\pi(\psi_{u_{ii}} \mid -) \sim \mathrm{Ga}\left(a_g + \frac{r}{2}, b_g + \frac{1}{2}\delta_i^T\delta_i\right).$$

9. The conditional posteriors for the model residuals of trait i, σ_i^{-2}, is:

$$\pi(\sigma_i^{-2} \mid -) \sim \mathrm{Ga}\left(a_r + \frac{n}{2}, b_r + \frac{1}{2}\sum_{j=1}^{n}\left(y_{ij} - \mathbf{x}^{(j)}\mathbf{b}_i - \mathbf{f}^{(j)}\lambda_i - \mathbf{z}^{(j)}\delta_i\right)^2\right).$$

Other random effects, such as the line $\times$ sex effects modeled in the gene expression example of this paper can be incorporated into this sampling scheme in much the same way as the residual genetic effects, $\boldsymbol{\Delta}$, are included here.

7 Acknowledgments

We would like to thank Barbara Englehardt and Iulian Pruteanu-Malinici for comments and advice on this method.

References

Armagan, A., D. B. Dunson, and M. Clyde, 2011 Generalized Beta Mixtures of Gaussians. Neural Information Processing Systems**:** arXiv:1107.4976.

Atkinson, B. and T. Therneau, 2012 kinship: mixed-effects Cox models, sparse matrices, and modeling data from large pedigrees. R package version 1.1.3.

Ayroles, J. F., M. A. Carbone, E. A. Stone, K. W. Jordan, R. F. Lyman, M. M. Magwire, S. M. Rollmann, L. H. Duncan, F. Lawrence, R. R. H. Anholt, and T. F. C. Mackay, 2009 Systems genetics of complex traits in Drosophila melanogaster. Nat Genet *41*(3)**:** 299–307.

Bhattacharya, A. and D. B. Dunson, 2011 Sparse Bayesian infinite factor models. Biometrika *98*(2)**:** 291–306.

Bickel, P. J. and E. Levina, 2008a Covariance regularization by thresholding. Annals of Statistics **36:** 2577–2604.

Bickel, P. J. and E. Levina, 2008b Regularized estimation of large covariance matrices. Annals of Statistics **36:** 199–227.

Biswas, S., J. D. Storey, and J. M. Akey, 2008 Mapping gene expression quantitative trait loci by singular value decomposition and independent component analysis. BMC Bioinformatics **9:** 244.

Cantor, R. M., K. Lange, and J. S. Sinsheimer, 2010 Prioritizing GWAS Results: A Review of Statistical Methods and Recommendations for Their Application. Am J Hum Genet *86*(1)**:** 6–22.

Carvalho, C. M., J. Chang, J. E. Lucas, J. R. Nevins, Q. Wang, and M. West, 2008 High-Dimensional Sparse Factor Modeling: Applications in Gene Expression Genomics. Journal of the American Statistical Association *103*(484)**:** 1438–1456.

Cheverud, J. M., 1996 Developmental Integration and the Evolution of Pleiotropy. Integrative and Comparative Biology *36*(1)**:** 44–50.

de la Cruz, O., X. Wen, B. Ke, M. Song, and D. L. Nicolae, 2010 Gene, region and pathway level analyses in whole-genome studies. Genet Epidemiol **34:** 222–231.

de Los Campos, G. and D. Gianola, 2007 Factor analysis models for structuring covariance matrices of additive genetic effects: a Bayesian implementation. Genet Sel Evol *39*(5)**:** 481–494.

el Karoui, N., 2008 Operator norm consistent estimation of large dimensional sparse covariance matrices. Annals of Statistics **36:** 2717–2756.

Engelhardt, B. E. and M. Stephens, 2010 Analysis of Population Structure: A Unifying Framework and Novel Methods Based on Sparse Factor Analysis. PLoS Genet *6*(9)**:** e1001117.

Fan, J., J. Lv, and L. Qi, 2011 Sparse High Dimensional Models in Economics. Annu Rev Econom **3:** 291–317.

Gibson, G. and B. Weir, 2005 The quantitative genetics of transcription. Trends Genet *21*(11)**:** 616–623.

Goldstein, H., 2010 *Multilevel Factor Analysis, Structural Equation and Mixture Models*, pp. 189–200. John Wiley & Sons, Ltd.

Hahn, P. R., C. M. Carvalho, and S. Mukherjee, 2013 Partial Factor Regression. Journal of the American Statistical Association.

Hansen, T. F. and D. Houle, 2008 Measuring and comparing evolvability and constraint in multivariate characters. J Evol Biol *21*(5): 1201–1219.

Henderson, C. R., 1984 *Applications of linear models in animal breeding*. Guelph, Ca: University of Guelph.

Hine, E. and M. W. Blows, 2006 Determining the effective dimensionality of the genetic variance-covariance matrix. Genetics *173*(2): 1135–1144.

Houle, D., 2010 Colloquium papers: Numbering the hairs on our heads: the shared challenge and promise of phenomics. Proc. Natl. Acad. Sci. U.S.A. **107 Suppl 1:** 1793–1799.

Huang, D. W., B. T. Sherman, and R. A. Lempicki, 2009a Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucl. Acids Res. *37*(1): 1–13.

Huang, D. W., B. T. Sherman, and R. A. Lempicki, 2009b Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc *4*(1): 44–57.

Kirkpatrick, M., 2009 Patterns of quantitative genetic variation in multiple dimensions. Genetica *136*(2): 271–284.

Kirkpatrick, M. and K. Meyer, 2004 Direct estimation of genetic principal components: Simplified analysis of complex phenotypes. Genetics *168*(4): 2295–2306.

Kruuk, L. E. B., 2004 Estimating genetic parameters in natural populations using the 'animal model'. Philos T Roy Soc B *359*(1446): 873–890.

Lande, R., 1979 QUANTITATIVE GENETIC-ANALYSIS OF MULTIVARIATE EVOLUTION, APPLIED TO BRAIN - BODY SIZE ALLOMETRY. Evolution *33*(1): 402–416.

Lynch, M. and B. Walsh, 1998 *Genetics and analysis of quantitative traits* (1 ed.). Sinauer Associates Inc.

Mcgraw, E. A., Y. H. Ye, B. Foley, S. F. Chenoweth, M. Higgie, E. Hine, and M. W. Blows, 2011 HIGH-DIMENSIONAL VARIANCE PARTITIONING REVEALS THE MODULAR GENETIC BASIS OF ADAPTIVE DIVERGENCE IN GENE EXPRESSION DURING REPRODUCTIVE CHARACTER DISPLACEMENT. Evolution *65*(11)**:** 3126–3137.

Meyer, K., 2007 WOMBAT—A tool for mixed model analyses in quantitative genetics by restricted maximum likelihood (REML). Journal of Zhejiang University-SCIENCE B *8*(11)**:** 815–821.

Meyer, K., 2009 Factor-analytic models for genotype x environment type problems and structured covariance matrices. Genet Sel Evol **41:** 21.

Meyer, K. and M. Kirkpatrick, 2010 Better estimates of genetic covariance matrices by "bending" using penalized maximum likelihood. Genetics *185*(3)**:** 1097–1110.

Mezey, J. and D. Houle, 2005 The dimensionality of genetic variation for wing shape in Drosophila melanogaster. Evolution *59*(5)**:** 1027–1038.

Morrissey, M. M., 2010 pedantics: Functions to facilitate power and sensitivity analyses for genetic studies of natrual popualtions.

Parts, L., O. Stegle, J. Winn, and R. Durbin, 2011 Joint genetic analysis of gene expression data with inferred cellular phenotypes. PLoS Genet *7*(1)**:** e1001276.

R Core Team, 2012 R: A Language and Environment for Statistical Computing. ISBN 3-900051-07-0.

Rausher, M. D., 1992 THE MEASUREMENT OF SELECTION ON QUANTITATIVE TRAITS - BIASES DUE TO ENVIRONMENTAL COVARIANCES BETWEEN TRAITS AND FITNESS. Evolution *46*(3)**:** 616–626.

Schluter, D., 1996 Adaptive radiation along genetic lines of least resistance. Evolution *50*(5)**:** 1766–1774.

Sorensen, D. and D. Gianola, 2010 *Likelihood, Bayesian and MCMC Methods in Quantitative Genetics (Statistics for Biology and Health)*. Springer.

Stegle, O., L. Parts, R. Durbin, and J. Winn, 2010 A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. PLoS Comput Biol *6*(5)**:** e1000770.

Stone, E. A. and J. F. Ayroles, 2009 Modulated modularity clustering as an exploratory tool for functional genomic inference. PLoS Genet *5*(5)**:** e1000479.

Walsh, B. and M. W. Blows, 2009 Abundant Genetic Variation plus Strong Selection = Multivariate Genetic Constraints: A Geometric View of Adaptation. Annu Rev Ecol Evol S **40:** 41–59.

Xiong, Q., N. Ancona, E. R. Hauser, S. Mukherjee, and T. S. Furey, 2012 Integrating genetic and gene expression evidence into genome-wide association analysis of gene sets. Genome Research **22:** 386–397.

Zhou, X. and M. Stephens, 2012 Genome-wide efficient mixed-model analysis for association studies. Nature Genetics **44:** 821–824.